Comparison of validity, repeatability and reproducibility of the Peer Assessment Rating (PAR) between digital and conventional study models

Sridhar Pasapula,[*] Martyn Sherriff,[†] Jeremy Breckon,[*] Dirk Bister[*] and Stefan Abela[*]

Guy's and St Thomas' NHS Foundation Trust[*] and King's College London,[†] London, UK

Introduction: The validity, reliability and inter-method agreement of Peer Assessment Scores (PAR) from acrylic models and their digital analogues were assessed.

Method: Ten models of different occlusions were digitised, using a 3 Shape R700 laser scanner (Copenhagen, Denmark). Each set of models was conventionally and digitally PAR-scored twice in random order by 10 examiners. The minimum time between repeat measurements was two weeks. The repeatability was assessed by applying Carstensen's analysis. Inter-method agreement (IEMA) was assessed by Carstensen's limit of agreement (LOA).

Results: Intra-examiner repeatability (IER) for the unweighted and weighted data was slightly better for the conventional rather than the digital models. There was a slightly higher negative bias of -1.62 for the weighted PAR data for the digital models. IEMA for the overall weighted data ranged from -8.70 – 5.45 (95% Confidence Interval, CI). Intra-class Correlation Coefficients (ICC) for the weighted data for conventional, individual and average scenarios were 0.955 (0.906 – 0.986 CI), 0.998 (0.995 – 0.999 CI). ICC for the weighted digital data, individual and average scenarios were 0.99 (0.97 – 1.00) and 1.00. The percentage reduction required to achieve an optimal occlusion increased by 0.4% for the digital scoring of the weighted data.

Conclusion: Digital PAR scores obtained from scanned plastic models were valid and reliable and, in this context, the digital semi-automated method can be used interchangeably with the conventional method of PAR scoring.

(Aust Orthod J 2016; 32: 184-192)

Received for publication: October 2015
Accepted: June 2016

Sridhar Pasapula: spasapula5@gmail.com; Martyn Sherriff: martyn.Sherriff@bristol.ac.uk; Jeremy Breckon: jeremy.breckon@kcl.ac.uk; Dirk Bister: dirk.bister@kcl.ac.uk; Stefan Abela: abelastefan@gmail.com

Introduction

The Peer Assessment Rating (PAR) and the digitisation of study models are now part of contemporary orthodontics and are routinely used to assess orthodontic treatment outcomes.

The PAR index is a validated epidemiological tool, originally developed to monitor orthodontic standards in the United Kingdom.[1] It measures pre- and post-treatment scores to determine the improvement of a malocclusion.[2]

The Orthodontic Society Clinical Standards Committee (2009) in the United Kingdom[3] published guidelines and advocated the use of the PAR index to assess treatment outcomes of patients. A high standard of practice dictates that a mean percentage reduction in PAR score should be high at more than 70%. The PAR index is widely used across Europe and has been shown to be valid and reliable[1,2,4,5] in its ability to quantify the extent of improvement and treatment success using plaster models.

Validity is defined as the extent to which a measure represents the object of interest. Accuracy is often used interchangeably.[6] Colton stated that accuracy encompasses a lack of bias, the tendency to arrive at a true value, precision, and the spread of a series of observations.[7] Roberts et al. defined validity as the

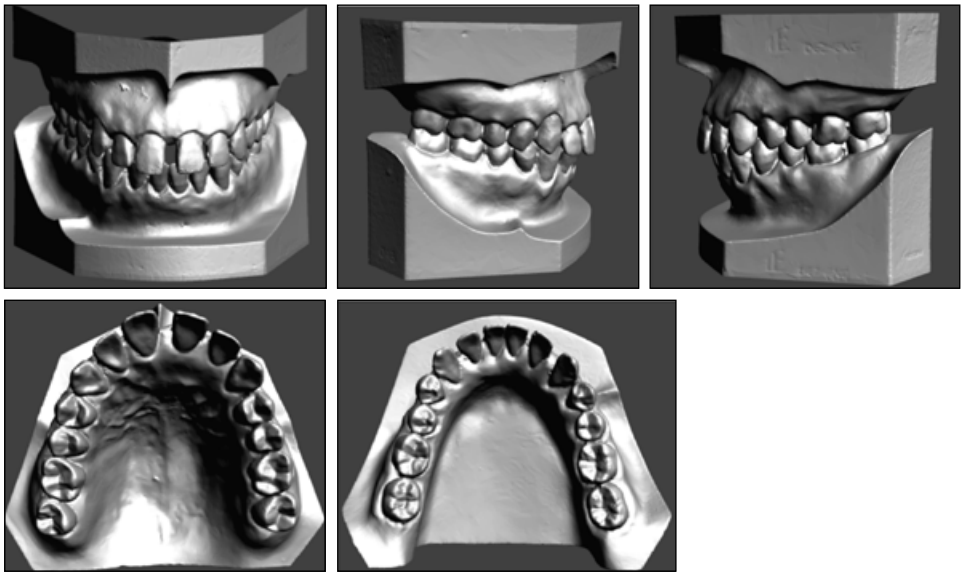

Figure 1. Digital models.

Figure 2. Resin models.

extent to which a measurement describes what it purports.[8]

Repeatability is defined by Nic et al.[9] as the closeness of agreement between independent results obtained with the same method on identical test material under identical conditions. Furthermore, Nic et al. defined reproducibility as the closeness of agreement between independent results obtained with the same method on identical test material, under different conditions.[9]

Reliability encompasses both repeatability and reproducibility.

The total PAR Index is measured by analysing the scores of 11 individual traits: upper right and left segments, upper and lower anterior segments, lower left and right segments, right and left buccal occlusion, overjet, overbite and centreline. The individual traits are weighted according to Richmond et al.,[2] resulting in the weighted PAR Index.

Three-dimensional (3D) imaging and the application of digital study models in contemporary orthodontic practice are also well documented. These applications include assessing orthodontic relationships of the dentition and facial form, 3D virtual treatment objectives and 3D custom-made archwires.[10]

The features of conventional plaster models include the ability to manipulate the models manually and ability to mount them on an articulator,[11] whilst the

benefits of digital models include a lack of physical storage, easy access and the ability to perform a diagnostic set-up.

The use of digital models should make no difference to a clinician's ability to diagnose and treatment plan,[12] which were findings confirmed by Whetten et al.[13]

Most previous studies have compared the assessment of digital against conventional models;[14] however, the present study is the first to investigate the plausibility of developing a computer-based PAR calibration program. Additional aims of this study included an investigation of the effect of multiple measurements on the method error and discrepancies in inter-examiner measurement.

To the authors' knowledge this is the first study utilising hardware and software components from the same manufacturer (3 Shape R700 Laser Scanner™ and 3 Shape ESM Software™, Copenhagen, Denmark).

All examiners were PAR-calibrated. The study objectives were:

1. To assess the intra-examiner repeatability of PAR scores for 10 acrylic and digital models;
2. To assess inter-examiner method agreement (IEMA) of PAR scores taken from 10 different occlusions;
3. To determine the difference in percentage reduction between the two methods; and
4. To assess the reliability of the PAR score of both weighted and unweighted components.

The *P* values for statistically significant differences were set at < 0.05 for:

1. intra-examiner measurements of digital and conventional models;
2. IEMA of PAR; and
3. Percentage PAR score reduction required to achieve a residual optimal PAR score of 2 points between the two methods.

The null hypotheses for the present study included the following:

- There is no clinically significant difference between intra-examiner measurements of digital models and conventional models ($p < 0.05$).
- There are no clinically significant differences in IEMA of PAR scores ($p < 0.05$).
- There is no clinically significant difference between the percentage PAR score reduction required to achieve a residual optimal PAR score of 2 points between the two methods ($p < 0.05$).

Materials and methods

The study used 10 Angle-based resin models (Smedent Medical Instrument Co, Ltd Shanghai, P.R. China), which encompassed the four incisor classifications based on the British Standards Institute (1983).[15] Nine models were of pretreatment occlusions and one model was classified as a post-treatment result.

The inclusion criterion for the dental models was a complete adult dentition from first molar to the contralateral first molar. The exclusion criteria included: hypodontia / supernumerary teeth; models of high standard with no voids, fractured teeth or any other damage; and no heavily restored teeth.

Ten PAR-calibrated examiners were recruited. The examiners scored the 10 conventional acrylic models twice followed by scoring all of the digital models, and adhered to this order throughout the entire study.

A .pdf tutorial based upon the proprietary 3 Shape ESM Software™ (Copenhagen, Denmark) was sent via electronic mail to the 10 examiners, a minimum of two weeks prior to the first digital scoring session to standardise the digitisation procedure.

The 10 Angle-based acrylic models were digitised using the 3 Shape R700 laser scanner and measurements were made on the 3D digital surface screen. A 20 inch, 32 bit colour LCD screen (Sony, Tokyo, Japan) was used to visualise the digital models with a resolution of 72bph 1280/1024 pixels.

The 10 models consisted of three Class I malocclusions, three Class II division 1 malocclusions, one Class II division 2 and three Class III malocclusions as determined by British Standard Institute Classification (1983).[15] PAR scoring was carried out in accordance with the Richmond protocol (1992).[1] The examiners independently scored the models and additional technical support with respect to the use of the ESM software was provided when necessary.

The digital and conventional models were scored using the Google 1–10 random number generator. If a model that had been previously scored was nominated again, the random number generator was re-looped until a model that had not previously been scored was chosen.

Table I. Unweighted and weighted PAR score Bias and Limits of Agreement (LOA).

Model number	Bias (PAR points)	LOA	ADRU (PAR points)
Overall score for unweighted data	-1.32	-6.53 – 3.89	10.42
			ADRW (PAR points)
Overall score for weighted data	-1.62	-8.70 – 5.45	14.15

Bias: conventional minus digital PAR scores.
LOA: Limits of agreement.
ADRU: Absolute difference for unweighted data.
ADRW: Absolute difference for weighted data.

Table II. Carstensen analysis for weighted PAR scores.

	Digital repeatability IER (DW)	Conventional repeatability IER (CW)	ADRW (PAR points)
Overall weighted repeatability	5.30	5.03	0.27
Model 1	6.45	6.45	0
Model 2	6.55	6.55	0
Model 3	7.66	5.79	1.87
Model 4	4.76	4.76	0
Model 5	5.32	3.39	1.93
Model 6	4.82	2.47	2.35
Model 7	2.90	2.90	0
Model 8	6.11	5.79	0.32
Model 9	0.50	0.50	0
Model 10	5.65	5.65	0

IER (DW): intra-examiner repeatability for digital weighted data.
IER (CW): intra-examiner repeatability for conventional weighted data.
ADRW: absolute difference for weighted points.

Table III. Carstensen analysis for unweighted PAR scores.

	Digital repeatability IER (DU)	Conventional Repeatability IER (CU)	ADRU (PAR points)
Overall unweighted repeatability	4.30	3.21	1.09
Model 1	5.72	2.84	2.88
Model 2	6.31	5.30	1.01
Model 3	3.82	3.77	0.05
Model 4	3.62	3.62	0
Model 5	4.83	2.05	2.78
Model 6	4.4	1.8	2.6
Model 7	2.79	2.79	0
Model 8	4.89	2.13	2.76
Model 9	0.50	0.50	0
Model 10	3.55	3.55	0

IER (DU): intra-examiner repeatability for digital unweighted data.
IER (CU): intra-examiner repeatability for conventional unweighted data.
ADRU: absolute difference in unweighted points.

The resin casts were measured with original PAR rulers. Digital models were scored using the proprietary software and its associated questionnaires (ESM Digital Solutions Ltd, Dublin, Ireland). Weighted PAR scores were determined using European component scores. To prevent memory recall a minimum of two weeks from initial scoring was used for repeats.

Data analysis

The data were entered into an Excel® 2013 spreadsheet (Microsoft Office 2013, IL, USA) and analysed using StataCorp. 2013 (Stata Statistical Software, TX, USA). Benedix Carstensen's analysis[16] for repeat measurements of the 10 models on the replicated PAR data was used to establish levels of agreement (LOA) between methods for each model for all operators for both unweighted and weighted PAR scores (Table I). Intra-examiner repeatability (IER) was determined using Carstensen's analysis for the conventional and digital data for both weighted (Table II) and unweighted PAR scores (Table III). Intra-class correlation coefficients (ICC) were used to assess the entire data set, each model independently

Table IV. ICC, PAR unweighted.

Variable	Method	ICC (CI 95) individual	ICC (CI 95) average
Maxillary Anterior	Conventional	0.956 (0.908-0.986)	0.998 (0.995-0.999)
Maxillary Anterior	Digital	0.971 (0.939-0.991)	0.999 (0.997-0.100)
Mandibular Anterior	Conventional	0.884 (0.777-0.962)	0.994 (0.986-0.998)
Mandibular Anterior	Digital	0.944 (0.886-0.982)	0.997 (0.994-0.999)
Buccal A-P	Conventional	0.524 (0.324-0.792)	0.959 (0.910-0.988)
Buccal A-P	Digital	0.578 (0.377-0.826)	0.966 (0.927-0.990)
Buccal Transverse	Conventional	0.673 (0.478-0.876)	0.977 (0.951-0.993)
Buccal Transverse	Digital	0.628 (0.428-0.853)	0.973 (0.940-0.992)
Buccal Vertical	Conventional	-	-
Buccal Vertical	Digital	-	-
Overjet	Conventional	0.979 (0.954-0.994)	0.999 (0.998-0.100)
Overjet	Digital	0.986 (0.970-0.996)	0.999 (0.999-1.000)
Overbite	Conventional	0.901 (0.807-0.969)	0.995 (0.989-0.998)
Overbite	Digital	0.928 (0.855-0.978)	0.996 (0.992-0.999)
Midline	Conventional	0.894 (0.794-0.966)	0.994 (0.988-0.998)
Midline	Digital	0.897 (0.798-0.967)	0.995 (0.988-0.998)
Overall	Conventional	0.955 (0.906-0.986)	0.998 (0.995-0.999)
Overall	Digital	0.980 (0.959-0.994)	0.999 (0.998-0.100)

CI 95: 95% Confidence interval.

Table V. ICC, PAR weighted.

Variable	Method	ICC (CI 95) individual	ICC (CI 95) average
Maxillary Anterior	Conventional	0.956 (0.908-0.986)	0.998 (0.995-0.999)
Maxillary Anterior	Digital	0.971 (0.939-0.991)	0.999 (0.997-1.00)
Mandibular Anterior	Conventional	0.884 (0.776-0.963)	0.994 (0.986-0.998)
Mandibular Anterior	Digital	0.944 (0.886-0.983)	0.997 (0.994-0.999)
Buccal A-P	Conventional	0.524 (0.324-0.792)	0.959 (0.910-0.988)
Buccal A-P	Digital	0.578 (0.377-0.826)	0.966 (0.927-0.990)
Buccal Transverse	Conventional	0.673 (0.478-0.876)	0.977 (0.951-0.993)
Buccal Transverse	Digital	0.628 (0.428-0.853)	0.973 (0.940-0.992)
Buccal Vertical	Conventional	-	-
Buccal Vertical	Digital	-	-
Overjet	Conventional	0.979 (0.954-0.994)	0.999 (0.998-1.00)
Overjet	Digital	0.986 (0.970-0.996)	0.999 (0.999-1.000)
Overbite	Conventional	0.901 (0.807-0.969)	0.995 (0.989-0.998)
Overbite	Digital	0.928 (0.855-0.978)	0.996 (0.992-0.999)
Midline	Conventional	0.894 (0.794-0.966)	0.994 (0.988-0.998)
Midline	Digital	0.897 (0.798-0.967)	0.995 (0.988-0.998)
Overall	Conventional	0.98 (0.95-0.99)	1.00 (1.00-1.00)
Overall	Digital	0.99 (0.97-1.00)	0.99 (0.97-1.00)

CI 95: 95% Confidence interval.

and to compare reliability for both unweighted (Table IV) and weighted data (Table V). The conventional and digital data were analysed with their respective weighted and unweighted components using a simple one-way random effects model for average and individual scenarios.

The recommended level for inter-examiner agreement was set at +/- 2 PAR points, which is the perceived level of clinical significance for pretreatment models.[17] For the single post-treatment model (Model 9), the following categories of orthodontic treatment results based on final PAR scores were identified in the literature: Acceptable < 5; Marginal 5–10; Poor > 10.[18] A maximum difference of +/- 5 points between the two methods was chosen for post-treatment study models. The threshold for acceptable ICC was set using reference data from medicine:[19] < 0.40 = poor, 0.40 – 0.59 = fair, 0.60 – 0.74 = good and > 0.74 = excellent. ICC Intra-class correlation coefficients should be above 0.8 for one method to be able to replace the other.

The univariate summary statistics were used to determine the mean PAR scores for all models. The percentage reduction required to achieve a residual PAR score of two points was calculated. The absolute differences between the methods were also calculated for each model.

Results

The difference in IER between methods describes the difference of how an individual interprets the PAR score for a model on repetition between methods. IER for unweighted scores on conventional models was 3.21 PAR points compared to digital models at 4.30 PAR points. The IER for weighted scores for conventional and digital models were 5.03 and 5.30 respectively.

The ideal score is 0; therefore, the lower the score, the better the repeatability. The repeatability for both unweighted and weighted scores were better for conventional than for digital scores: 3.21 and 5.03 for conventional models versus 4.30 and 5.30 for digital models, respectively.

The difference in IER between conventional and digital models for unweighted and weighted scores was referred to as an absolute difference in repeatability for unweighted data (ADRU) and an absolute difference in repeatability for weighted data (ADRW). Greater differences were shown in the inter-method IER for unweighted rather than weighted data; overall ADRU was 1.09 and overall ADRW was 0.27.

In the present study, bias was assessed as the difference in PAR scores between conventional and digital models; a negative bias caused the digital PAR scores to be higher. The digital models had a tendency to over score and hence had a negative bias of -1.32 and -1.62 PAR points for both unweighted and weighted PAR data.

The IEMA is the difference in interpretation of PAR scores for a model measured using the conventional and the digital techniques. The overall IEMA was 10.42 PAR points for unweighted data and 14.15 PAR points for weighted data.

The ICCs were calculated using a One Way Random Effects Model for Absolute Agreement and the overall ICC for conventional and digital data was 0.955 and 0.980. This was much higher than the threshold for a good ICC of 0.8. Individual components that were least reliable were anterior, posterior and transverse buccal occlusion. Centre lines and overbite were lower than overall ICC; however, still above the threshold of 0.8.

The ICC reliability scores for physical models were lowest for the lower anterior segment contact points at 0.884 CI 95 (0.777 – 0.962), albeit still within acceptable limits.[20]

Discussion

The advantages of utilising 3D study models in everyday practice have been well documented and have been previously investigated for validity and reliability.[21]

To the authors' knowledge all existing studies evaluated different aspects of the 3D digital models' practicality and ease of use; in particular, software programs in combination with hardware from different manufacturers. Tomassetti et al.[22] and Mullen et al.[23] investigated the precision of the Bolton tooth-size discrepancy whilst other authors compared different software program used to analyse metric measurements of digital models.[24]

The present study is the first to use hardware and software components from the same manufacturer: 3 Shape R700™ Laser Scanner (Copenhagen, Denmark) and OrthoAnalyzer™3 Shape ESM

Software. Overall the findings are similar to those of Mayers et al.[14] and Stevens et al.,[25] all of which indicate that the reliability of PAR measurements between digital and conventional study models is acceptable.

The IER scores, which ideally should be as close to 0 as possible, were better on conventional models for both weighted and unweighted data. This was most likely because the overjet was scored inconsistently for the conventional models and led to the intra-examiner repeatability decreasing less for the digital models; the scoring inconsistency led to a favourable decrease only in the case of conventional models.

Greater differences were shown in the inter-method IER for unweighted (ADRU) rather than weighted data (ADRW); 1.09 compared with 0.27. It would be expected that there are fewer differences in the inter-method IER with unweighted data; however, the findings of the present study showed the opposite. The majority of inconsistent scoring in IER was associated with the unweighted aspects of the PAR index; therefore, when the weighted aspects are taken into account, the overall score diminished, improving the overall ADRW value. The weighted components may have been scored at different levels of inconsistency between the methods for an individual component or multiple components (e.g., an increase in centreline score but a reduction in overbite) and this may have led to a greater deterioration / improvement in repeatability of a model. Differences in the scoring of weighted elements may have cancelled each other out, leaving most of the variability between the unweighted components such as buccal occlusion, as confirmed with the inconsistent scoring from the ICC data.

The negative bias associated with digital models for the unweighted and weighted data was due to over-scoring, which could be explained by the limitations of the software which overestimated contact point displacement by adding a vertical component to the horizontal measurement. An analogy in algebraic terms would be the measurement of the hypotenuse instead of the adjacent side. The slight increase in negative bias for weighted data could be due to physical limitations of the examiners to accurately analyse the overjet, as in the case of crowding of the lower labial segment.

Mayers et al.[14] had shown that the least reliable component of the PAR index is the buccal occlusion; however, the present study highlights the difficulties of measuring displaced teeth.

The overall IEMA was worse for weighted data compared with unweighted data by a total of 3.73 PAR points.

It is plausible that a lack of agreement of the scores between examiners may have accounted for the majority of the disagreement in the unweighted data. This possibly led to an increase in disagreement between the two methods when the weighted data were analysed, and hence amplified the disagreement in the unweighted data. The poorer overall IEMA for weighted data was likely because the weighted aspects of the PAR index were scored more inconsistently by the examiners using the two methods, particularly for borderline measurements, and this may have been amplified in the weighted data.

In the current investigation, all aspects of the PAR index associated with linear measurements such as contact point displacements and overjet scored sufficiently high, above 0.8. The least reliable linear measurement was lower anterior segment crowding, and this could be explained by the less stringent way the examiners scored the contact points which are less heavily weighted.

Other parameters such as overbite and buccal occlusion are more subjectively assessed and based mainly on visual interpretation. They cannot be easily quantified either conventionally with the PAR ruler, or with the PAR scoring software, which does not have the functionality.

The zoom feature provided by the investigated software has the potential to amplify irregularities in the buccal occlusion, making consistent scoring difficult for the examiners. Although the overjet was quantitatively measured, several examiners noted difficulty in using this feature of the software.

Centerline deviation, overbite depth and antero-posterior buccal occlusion are visually assessed in two dimensions, although three-dimensional representation is available in the form of conventional models and 'pseudo' three-dimensional representation in the form of virtual models. Crossbites and scissors bites would benefit from a full three-dimensional assessment.

Malik et al.[26] investigated whether medico-legal information from study models could be obtained from post-debond photographs. It was determined that this was viable and reliable for all parameters except the overbite. The digital ICC was slightly higher

for the conventional models. This could be due to the software that allowed the examiners to make the models translucent to enable a determination of the depth of overbite, which is not possible with the conventional models.

Scoring the 'ideal occlusion' of Model 9 could have statistically weakened the results; however, using this model was crucial to having a 'gold standard' to which the other models could be compared and scored.

One of the strengths of the present study was the number of repeat measurements that multiple examiners were able to perform on both the conventional and digitised models. The original Bland–Altman method was developed for two sets of measurements done on one occasion (i.e. independent data), and so this approach was not suitable for repeated measures of data.[27] However, the Carstensen analysis[16] allows for repeat data and uses linear regression to calculate predictive equations. This provided the rationale to determine LOA for IEMA and IER for conventional and digital PAR data.

In addition, all examiners were PAR-calibrated, unlike a previous study,[28] and this increased the robustness of the measured data. Based on the findings of the present study, digital PAR scores were valid and reliable and clinicians may safely utilise digital models and their PAR index.

Conclusions

- Negative bias measuring the PAR score difference between digital and conventional models was not clinically significant.
- The IER was slightly better for conventional models than digital models for unweighted and weighted data.
- The LOAs were narrow enough to allow one method to replace the other.
- Overall ICC scores for conventional and digital models for unweighted and weighted data were above a threshold of 0.8, suggesting sufficient agreement for one method to replace another.
- The digital PAR scores were deemed to be sufficiently valid and reliable to be used interchangeably with conventional PAR scores.

Acknowledgments

The authors would like to thank examiners: Huw Jeremiah, Inas Naser, Mandeep Gosal, Sally Zahran, Andrew Noon, Nickolas Maschas, and Victoria Johnson for participating in this study and recording the PAR values for the dental models provided.

Corresponding author

Stefan Abela
Department of Orthodontics
Guy's and St Thomas' NHS Foundation Trust
Floor 25, Tower Wing
Great Maze Pond
London
SE1 9RT

Email: stefan.abela@gstt.nhs.uk

References

1. Richmond S, Shaw WC, O'Brien KD, Buchanan IB, Jones R, Stephens CD et al. The development of the PAR Index (Peer Assessment Rating): reliability and validity. Eur J Orthod 1992;14:125-39.
2. Richmond S, Shaw WC, Roberts CT, Andrews M. The PAR Index (Peer Assessment Rating): methods to determine outcome of orthodontic treatment in terms of improvement and standards. Eur J Orthod 1992;14:180-7.
3. Committee BOSCS. Guidelines for Primary Care Trusts and Local Health Boards to assess the treatment outcome of patients treated by specialist orthodontists or dentists using the Peer Assessment Rating (PAR) Index. 2009.
4. Firestone AR, Beck FM, Beglin FM, Vig KW. Evaluation of the peer assessment rating (PAR) index as an index of orthodontic treatment need. Am J Orthod Dentofacial Orthop 2002;122:463-9.
5. DeGuzman L, Bahiraei D, Vig KW, Vig PS, Weyant RJ, O'Brien K. The validation of the Peer Assessment Rating index for malocclusion severity and treatment difficulty. Am J Orthod Den-tofacial Orthop 1995;107:172-6.
6. Houston WJ. The analysis of errors in orthodontic measurements. Am J Orthod 1983;83:382-90.
7. Colton T. Controlled clinical trials. Am Rev Respir Dis 1974;110 Part 2:20-4.
8. Roberts CT, Richmond S. The design and analysis of reliability studies for the use of epidemiological and audit indices in orthodontics. Br J Orthod 1997;24:139-47.
9. M N. IUPAC Compendium of Chemical Terminology 2nd Edition. IUPAC Compendium of Chemical Terminology Gold Book, 2. 1997 (2nd Edition).
10. Hajeer MY, Millett DT, Ayoub AF, Siebert JP. Applications of 3D imaging in orthodontics: part I. J Orthod 2004;31:62-70.
11. Shastry S, Park JH. Evaluation of the use of digital study models in postgraduate orthodontic programs in the United States and Canada. Angle Orthod 2014;84:62-7.
12. Rheude B, Sadowsky PL, Ferriera A, Jacobson A. An evaluation of the use of digital study models in orthodontic diagnosis and treatment planning. Angle Orthod 2005;75:300-4.
13. Whetten JL, Williamson PC, Heo G, Varnhagen C, Major PW. Variations in orthodontic treatment planning decisions of Class II

patients between virtual 3-dimensional models and traditional plaster study models. Am J Orthod Dentofacial Orthop 2006;130:485-91.
14. Mayers M, Firestone AR, Rashid R, Vig KW. Comparison of peer assessment rating (PAR) index scores of plaster and computer-based digital models. Am J Orthod Dentofacial Orthop 2005;128:431-4.
15. Institution BS. British Standards Glossary of Dental Terms BS - 4492 London: BSI; 1983.
16. Carstensen B. Comparing and predicting between several methods of measurement. Biostatistics 2004;5:399-413.
17. Brown R, Richmond S. An update on the analysis of agreement for orthodontic indices. Eur J Orthod 2005;27:286-91.
18. King GJ, McGorray SP, Wheeler TT, Dolce C, Taylor M. Comparison of peer assessment rat-ings (PAR) from 1-phase and 2-phase treatment protocols for Class II malocclusions. Am J Or-thod Dentofacial Orthop 2003;123:489-96.
19. Cicchetti DV, Sparrow SA. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. Am J Ment Defic 1981;86:127-37.
20. Fleiss JL. Statistical Methods for Rates and Proportions, 2nd Edition. New York: Miley; 1981.
21. Bell A, Ayoub AF, Siebert P. Assessment of the accuracy of a three-dimensional imaging sys-tem for archiving dental study models. J Orthod 2003;30:219-23.
22. Tomassetti JJ, Taloumis LJ, Denny JM, Fischer JR Jr. A comparison of 3 computerized Bolton tooth-size analyses with a commonly used method. Angle Orthod 2001;71:351-7.
23. Mullen SR, Martin CA, Ngan P, Gladwin M. Accuracy of space analysis with emodels and plaster models. Am J Orthod Dentofacial Orthop 2007;132:346-52.
24. Santoro M, Galkin S, Teredesai M, Nicolay OF, Cangialosi TJ. Comparison of measurements made on digital and plaster models. Am J Orthod Dentofacial Orthop 2003;124:101-5.
25. Stevens DR, Flores-Mir C, Nebbe B, Raboud DW, Heo G, Major PW. Validity, reliability, and reproducibility of plaster vs digital study models: comparison of peer assessment rating and Bolton analysis and their constituent measurements. Am J Orthod Dentofacial Orthop 2006;129:794-803.
26. Malik OH, Abdi-Oskouei M, Mandall NA. An alternative to study model storage. Eur J Orthod 2009;31:156-9.
27. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.
28. Andrews CK. Validity and reliability of peer assessment rating index scores of digital and plaster models. Thesis. The Ohio State University, 2008.